DOCUMENT RESUME

ED 079 334 TM 002 932

AUTHOR	Blumenfeld, Warren S.; And Others
TITLE	Application of a New Psychometric Method to an Old Tried and Untrue Experimental Design to Improve the Validity of a Tailor-Made Scoring Key.
PUB DATE	73
NOTE	10p.; Paper presented at the American Personnel and Guidance Association Meeting (San Diego, California, February 1973)
JOURNAL CIT	San Diego Convention Abstracts; p142-143 1973
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Answer Keys; *Item Analysis; *Psychometrics; *Scoring; Speeches; *Validity
IDENTIFIERS	*Cross validation

ABSTRACT

This is a psychometric hoax paper, the purpose of which is to indicate once again the importance of cross-validation, particularly in the development of specially-keyed inventories. The junior author and the new psychometric method play critical roles in the study. Appropriate credit and references are present. (Author)

ED 079334

Application of a New Psychometric Method to an Old Tried and Untrue Experimental Design to Improve the Validity of a Tailor-Made Scoring Key[1]

Warren S. Blumenfeld and William L. Godbey

Georgia State University

and

Joshua C. Blumenfeld

Cliff Valley Nursery School, Atlanta, Georgia

U S DEPARTMENT OF HEALTH EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 002 932

It is difficult to know where to begin with a study as really special and exciting as this. As suggested by the title, the purpose of the study was to improve the validity of a tailor-made scoring key by the application of a new psychometric method to an old tried and untrue experimental design. Let me then first summarize (1) the background and (2) the procedure and results of the study, before getting into the details of the procedure--particularly the new psychometric method.

Background

There has developed over the past several years a body of literature in applied psychometrics that would indicate that the empirical development of tailor-made scoring keys is to be preferred to "store-bought" and/or a priori scoring keys. Further, the apparent plateau, or perhaps ceiling, for validity coefficients also seems to suggest the pressing need for breakthroughs in new and innovative psychometric methods and instruments.

However, as Kurtz pointed out so well in 1948, too often the wishes and hopes of the practitioner/developer and/or the consumer manifest themselves in a strange form of selective perception and self-deception in the evaluation of

FILMED FROM BEST AVAILABLE COPY

the effectiveness of such tailor-made keys, i.e., the acceptance of self-fulfilling "research" via the foldback design (that old tried and untrue design in which the tailor-made key is "tested" by re-applying it to the same data base from which it was originally developed).

Procedure and Results

Regarding the current study, in the data collection phase, 100 special subjects responded to a special instrument (100 2-alternative items) through a special response mode. A new psychometric method was employed extensively in the data collection. Then, an equally special external criterion was developed with which the tailor-made key was subsequently developed.

Following data collection, there was accomplished an item analysis utilizing the special external criterion. The item analysis identified 24 of the 100 items for the special tailor-made key.

Application of this key in the same data base resulted in a biserial correlation of .99+. At this point the authors were extremely encouraged, as one might well imagine, both in terms of the new psychometric method and in terms of the key and the instrument.

However, it was decided to conduct the "academic nicety" of cross-validation. Application of the key in cross-validation resulted in a disappointing biserial correlation of .19.

The first coefficient reported is (clearly) significant beyond the .01 level; the second coefficient reported is not significant at the .05 level. Why the discrepancy? How account for the shrinkage? Or better, the inflation?

With this overview, perhaps we can retrace the research methodology so as to explain and better understand this discrepancy.

Purpose

Although the implied purpose of this study was to apply a new psychometric method to an old tried and untrue experimental design to improve the validity of a tailor-made scoring key, the real purpose was to point out (yes, once again) the specious, self-serving, insidious, suspect, spurious, fallacious, but facinating results that are obtained when cross-validation does not follow item analysis.

Procedure

With this last revelation in mind, let us now examine in more detail the procedure generally, and the new psychometric method specifically. And, unlike the traditional "senior author," let me give credit where credit is due regarding the relative contributions of the senior author, the second author, and, perhaps particularly, the junior author.

Data Collection

The subjects, the instrument (the new psychometric method), and the external criterion follow. As previously indicated, the subjects, the instrument, and the criterion were all quite special.

Subjects. Indeed the subjects were special; in fact, they did not exist except in the rich (however bizarre) imaginations of the authors. They are purely hypothetical, science fictional if you will. If credit must be taken, the senior author assumes the credit for the special subjects.

Instrument (The New Psychometric Method). The special instrument I now hold in my hand. As you can see, it is a United States penny, circa 1971. (Not being much of a grantsman, the project was run on an extremely modest and limited budget.) You will note that the coin has two sides, i.e., two alternatives. A flip of the instrument by the junior author established the

convention as to whether heads would be alternative A or B. (As it turned out, heads was B.) The second author then laid the "instrument" upon his thumb and proceeded to flip the coin 100 times for each of the 100 hypothetical subjects. If the coin came up heads, a B response was recorded; if the coin came up tails, an A response was recorded. The University generously provided the 100 answer sheets (the 100 subjects).

This coin flipping then was the "new psychometric method." As I will give an appropriate name to the new psychometric method later in the paper, I will at this juncture (demonstrating unusual self-discipline) resist the temptation of describing the method as "a series of one-tailed tests," or "the use of a digital computer," or even "one of cumulative side effects."

In this manner, that is, with this special instrument, and its attendant new psychometric method, 100 2-alternative responses were generated for each of the 100 special subjects. Let us now turn to the special external criterion used in the study.

Criterion. Following the development of the 100 100-response answer sheets, a step-wise algorithm was used to develop the special external criterion Specifically, the senior author (in the pedagogical spirit and tradition often suggested by students) stood at the top of an (outside) staircase and allowed the 100 answer sheets to tumble and float to the base of the staircase, littering the various individual stairs in the process. At this point, the junior author (eager to please, as junior authors are wont) recouped the answer sheets in whatever order (i.e., random) they had happened to fall. (I suppose one could refer to this as a "least-stairs solution.")

At this point, the second author stratified the 100 answer sheets into 2 stacks of 50 each, i.e., ne sorted, odd-even. Then the junior author flipped

another penny (the other half of the budget) to establish which stack of 50 would be the high criterion group and which stack of 50 would be the low criterion group. Following this re-application of the new psychometric method, in a blaze of scientific rigor, the junior author once again flipped the coin to establish which half of the high group and which half of the low group would be the item analysis (primary) group and which half of each criterion group would be the cross-validation (holdout) group (yet another application of the new psychometric method!) In this manner, 4 sets of 25 answer sheets, i.e., high primary, low primary, high holdout, and low holdout were established for the study. (We had originally planned to use several random tables in the criterion development phase. If we had, I suppose I could have now referred to these tools as "a number of random tables.")

Summary of data collection. Through the application of a "new psychometric method" (coin flipping), a data base of 100 2-alternative responses was generated for 100 (hypothetical) subjects. Utilizing a similarly generated special external criterion, these 100 answer sheets were further sub-divided into 4 analysis groups as per any traditional item analysis project.

Data Analysis

There were three phases to the data analysis of this research, i.e., (1) item analysis, (2) foldback, and (3) cross-validation

Item Analysis. The 100 items in the item pool were item analyzed using the procedure described by Lawshe and Baker (1950) with the special external criterion as previously described. In the item analysis, there were 25 in the high group and 25 in the low group. Alpha of .05 was used to identify the "discriminating" items for inclusion in the "special" key.

Foldback. The "items" surviving the item analysis were (re)applied to the answer sheets of the item analysis group. The predictive validity of the key was documented by biserial correlation.

Cross-Validation. However, for those more interested in the better (rather than the more fulfilling) estimate of the relationship between the derived key and the external criterion, the items surviving the item analysis were scored in the holdout groups of 25 high answer sheets and 25 low answer sheets. Again, biserial correlation was obtained to quantify the relationship between the special key and the criterion, i.e., the predictive validity.

Results

Item analysis procedure identified 24 items (chance would have been 5) which discriminated between the high and low groups at or beyond the .05 level. No doubt you will be interested in which items "came through." They were items 7, 10, 12, 16, 20, 21, 27, 31, 34, 35, 41, 42, 43, 59, 64, 66, 68, 72, 77, 81, 84, 88, 91, and 92. These item numbers are as meaningful in context as they are out of context (or vice versa?).

Applying these 24 items back upon the original sample in which they were derived, the obtained biserial correlation was .99+. No doubt, rounding error prevented the completely self-fulfilling prophecy. This was most encouraging, as this obtained coefficient is clearly off zero beyond the .05 level. (Consider here for a moment those of your acquaintance and/or your employ using this foldback design and at this point mouthing such quasi-professional, but sage, things as "Of course, these results should be interpreted with some caution.")

Unfortunately, when the 24-item key was applied to the holdout sample of 50, the encouraging coefficient of .99+ shrank slightly. In fact, it shrank back to .19 (not significantly off zero at the .05 level). Too bad; we felt we were on to something--both in terms of a new psychometric method and in terms of the operational utility of the key and instrument.

For those of you who are psychometric purists, you will be excited to learn that the obtained odd-even, corrected reliability of the key was .29 (N = 100).

Discussion and Conclusion

At this point, the reason for the obtained discrepancy between the foldback results and the cross-validation (hopefully) should be perfectly clear. The whole thing was a hoax; the old tried and untrue design, i.e. foldback, really did (and does) make something out of nothing--in this case out of something slightly less than nothing. Little or no further discussion seems necessary. In a sense (no pun intended) Cureton's classic paper (1952) has been re-executed. At the suggestion of the junior author (still eager to help), I call your attention to the recent treatment of this subject by the senior author (Blumenfeld, 1972). It seems (perhaps cruelly) clear once again that (1) the application of the key to the control group is the acid test of the quality of the key and (2) the (re)application of the key to the original group is but a half-acid test of the quality of the key.

One would think that this point has been well made often enough, but as an applied psychologist dealing with students and practitioners of business administration and/or educational administration, it is painfully clear to me

that the foldback design still remains very much in vogue. (For a recent insidious execution of the foldback experimental design, see, for example, Novak, 1970.) It is for that reason that I continue to believe that it is appropriate to beat home the point of cross-validation, i.e., let's have no more of this half-acid research!

Oh yes, there seem to be two pieces of business yet to be handled. These concern the junior author and the naming of the new psychometric method.

Regarding the junior author, he is now 5¼ years old. At the time of the study he was 3¾ years old. (The publication lag takes its toll on all of us.)

Regarding the naming of the new psychometric method, you will recall, that the explicit operational mechanics of the procedure were to lay the coin upon one's thumb and flip. Considering the non-consistency between the flippings (i.e., the application of the new psychometric method) of the second and junior authors, and, if you will not think it too flippant of me, I consider it uncommonly and punishingly appropriate to call the new psychometric method:

"THE METHOD OF NON-CONSTANT THUMBS"

And, at the risk of losing our place in psychometric history, the future application of this method is not recommended.

References

Blumenfeld, W. S. Quasi-successful concurrent validation of a special key for a relatively new and exciting personality instrument in a group of potential managers: or, I am never startled by a fish. Paper read at the meeting of the Georgia Psychological Association, Macon, May 1972. (Republished: Atlanta Economic Review, 1972, 22 (5), 14-15. Republished: The Industrial Psychologist, 1972, 9 (2) 23-26. Republished: The American Psychological Association Monitor, 1972, 3 (9,10), 3, 14. Republished: The Journal of Irreproducible Results, in press)

Cureton, E. E. Reliability, validity, and baloney. Educational and Psychological Measurement. 1950, 10, 94-96.

Kurtz, A. K. A research test for the Rorschach test. Personnel Psychology, 1948, 1, 41-51.

Lawshe, C. H., & Baker, P. C. Three aids in the evaluation of the significance of the difference between percentages. Educational and Psychological Measurement, 1950, 10, 263-270.

Novak, S. R. Developing an effective application blank. Personnel Journal 1970, May, 419-423.

Footnote

[1]Paper read at the meeting of the American Personnel and Guidance Association, San Diego, February 1973.